CONTENTS

6 Mixed Model Analysis

Basic model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where

$\mathbf{X}$ is a $n \times p$ model matrix of known constants

$\boldsymbol{\beta}$ is a $p \times 1$ vector of "fixed" unknown parameter values

$\mathbf{Z}$ is a $n \times q$ model matrix of known constants

$\mathbf{u}$ is a $q \times 1$ random vector

$\mathbf{e}$ is a $n \times 1$ vector of random errors

with

$$E(\mathbf{e}) = \mathbf{0} \qquad V(\mathbf{e}) = \mathbf{R}$$

$$E(\mathbf{u}) = \mathbf{0} \qquad V(\mathbf{u}) = \mathbf{M}$$

$$Cov(\mathbf{e}, \mathbf{u}) = 0$$

Then
$E(\mathbf{y}) =$
$V(\mathbf{y}) =$

6.1 Normal-theory mixed model

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right)$$

Then, $\mathbf{y} \sim$

Example 1. Random Blocks
Comparison of four processes for producing penicillin

$$\left.\begin{array}{l} Process\ A \\ Process\ B \\ Process\ C \\ Process\ D \end{array}\right\} \quad \begin{array}{l} \text{Levels of a "fixed"} \\ \text{treatment factor} \end{array}$$

Blocks correspond to different batches of an important raw material, corn steep liquor

- Random sample of five batches
- Split each batch into four parts:
  - run each process on one part
  - randomize the order in which the processes are run within each batch

Here, batch effects are considered as *random* block effects.

- Batches are sampled from a population of many possible batches
- To repeat this experiment you would need to use a different set of batches of raw material

Model:

$y_{ij} =$	$\mu + \alpha_i$	$+\beta_j$	$+e_{ij}$	
↑	↑	↑	↑	
Yield for the i-th process applied to the j-th batch	mean yield for the i-th process, averaging across the entire population of possible batches	random batch effect	random error	where

$$\begin{aligned} \beta_j &\sim NID(0, \sigma^2_\beta) \\ e_{ij} &\sim NID(0, \sigma^2_e) \end{aligned}$$

and any e_{ij} is independent of any β_j. Here $\mu_i =$

Restrict model with $\alpha_4 = 0$. Then

- $\mu =$
- $\alpha_i =$

In R we could use the "treatment" constraints where $\alpha_1 = 0$. Then

- $\mu =$
- $\alpha_i =$

Alternatively, we could choose the solution to the normal equations given by "sum" constraints.

- $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 =$
- $\mu =$
- $\alpha_i =$

Variance-covariance structure:

- $V(Y_{ij}) =$
- Different runs on the same batch: $Cov(Y_{ij}, Y_{kj}) =$
- Correlation among yields for runs on the same batch:$\rho =$

- Results for runs on different batches are uncorrelated (independent):

$$Cov(Y_{ij}, Y_{k\ell}) = 0 \quad \text{for j} \neq \ell$$

- Results from the four runs on a single batch:

This special type of covariance structure is called *compound symmetry*

Write this model as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e}$

Here
$\mathbf{M} =$
$\mathbf{R} =$

$V(\mathbf{y}) =$

Example 2. Hierarchical Random Effects Model
Analysis of sources of variation in a process used to monitor the production of a pigment paste.
Current Procedure:

- Sample barrels of pigment paste
- One sample from each barrel
- Send the sample to a lab for determination of moisture content

Measured Response: (Y) moisture content of the pigment paste (units of one tenth of 1%).
Problem: Variation in moisture content is too large

- average moisture content is approximately 25 (or 2.5%)
- standard deviation of about 6

Examine sources of variation:
Data Collection: Hierarchical (or nested) Study Design

- Sample b barrels of pigment paste
- s samples are taken from the content of each barrel
- Each sample is mixed and divided into r parts. Each part is sent to the lab.

There are
$n = (b)(s)(r)$ observations.

Model:

$$y_{ijk} = \mu + \beta_i + \delta_{ij} + e_{ijk}$$

where

- y_{ijk} is the moisture content determination for the k-th part of the j-th sample from the i-th barrel
- μ is the mean moisture content
- β_i is a random barrel effect:
$$\beta_i \sim NID(0, \sigma^2_\beta)$$
- δ_{ij} is a random sample effect:
$$\delta_{ij} \sim NID(0, \sigma^2_\delta)$$
- e_{ijk} corresponds to random measurement error:
$$e_ijk \sim NID(0, \sigma^2_e)$$

Covariance Structure:

- Homogeneous variances: $V(Y_{ijk}) =$

- Two parts of one sample:$Cov(Y_{ijk}, Y_{ij\ell}) =$
- Observations on different samples taken from the same barrel: $Cov(Y_{ijk}, Y_{im\ell}) =$
- Observations from different barrels:$Cov(Y_{ijk}, Y_{cm\ell}) =$

In this study

- $b = 15$ barrels were sampled
- $s = 2$ samples were taken from each barrel
- $r = 2$ sub-samples were analyzed from each sample taken from each barrel

Write this model in the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e}$$

where
$\mathbf{R} =$
$\mathbf{M} =$
Then
$E(\mathbf{y}) =$
$V(\mathbf{y}) =$

Example 3.

Researchers investigated the effects of 12 different drugs on the level of a protein in the blood of mice. On each of 5 days, 12 mice were randomly assigned to the 12 drugs with one mouse for each drug. Each mouse was injected with its assigned drug, and then blood samples were taken from each mouse at 4 time points: 1, 2, 3, and 4 hours after injection. The same process was repeated each day with 12 different mice, so a total of 60 mice were used in the experiment. The level of the protein of interest was measured in each of the 240 blood samples. For $i = 1, \ldots, 5$, $j = 1, \ldots, 12$, and $k = 1, \ldots, 4$, let y_{ijk} be the protein level measurement on day i for chemical drug j at time k. For $i = 1, \ldots, 5$, $j = 1, \ldots, 12$, and $k = 1, \ldots, 4$, consider the model

$$y_{ijk} = \mu_{jk} + d_i + e_{ijk},$$

where μ_{jk} terms are unknown fixed parameters and the other terms are random effects defined as follows. Let $\mathbf{d} = [d_1, \ldots, d_5]^T$. For $i = 1, \ldots, 5$ and $j = 1, \ldots, 12$, let $\mathbf{e}_{ij} = [e_{ij1}, \ldots, e_{ij4}]^T$.
Suppose

$$\mathbf{d} \sim N(\mathbf{0}, \sigma_d^2 \mathbf{I}_{5\times 5}),$$

and

$$\mathbf{e}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_e) \text{ for } i = 1, \ldots, 5 \text{ and } j = 1, \ldots, 12,$$

where σ_d^2 is an unknown positive variance parameter and

$$\boldsymbol{\Sigma}_e = \begin{bmatrix} \sigma_1^2 & \rho^4\sigma_1\sigma_2 & \rho^9\sigma_1\sigma_3 & \rho^{19}\sigma_1\sigma_4 \\ \rho^4\sigma_1\sigma_2 & \sigma_2^2 & \rho^5\sigma_2\sigma_3 & \rho^{15}\sigma_2\sigma_4 \\ \rho^9\sigma_1\sigma_3 & \rho^5\sigma_2\sigma_3 & \sigma_3^2 & \rho^{10}\sigma_3\sigma_4 \\ \rho^{19}\sigma_1\sigma_4 & \rho^{15}\sigma_2\sigma_4 & \rho^{10}\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}$$

for some unknown standard deviation parameter $\sigma_i > 0, i = 1, 2, 3, 4$ and some unknown correlation parameter ρ. Finally, suppose that $\mathbf{d}$ and $\mathbf{e}_{11}, \ldots, \mathbf{e}_{5,12}$ are all independent. In terms of model parameters, give a simplified expression for the variance of the generalized least squares estimator of each of the following:

(a) μ_{43}

(b) $\bar{\mu}_{4\cdot}$

(c) $\mu_{14} - \mu_{24}$

(d) $\mu_{41} - \mu_{44}$

6.2 Analysis of Mixed Linear Models

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where $\mathbf{X}_{n\times p}$ and $\mathbf{Z}_{n\times q}$ are known model matrices and

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{M} & 0 \\ 0 & \mathbf{R} \end{bmatrix} \right)$$

Then

$$Y \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma} = \mathbf{Z}\mathbf{M}\mathbf{Z}^T + \mathbf{R}$$

Some objectives

(i) Inferences about estimable functions of fixed effects: Point estimates, Confidence intervals and Tests of hypotheses.

(ii) Estimation of variance components (elements of $\mathbf{M}$ and $\mathbf{R}$)

(iii) Predictions of random effects (blup)

(iv) Predictions of future observations

6.3 Generalized Least Squares (GLS) Estimation

Suppose

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$

and also suppose

$$\boldsymbol{\Sigma} = V(\mathbf{y}) = \mathbf{Z}\mathbf{M}\mathbf{Z}^T + \mathbf{R}$$

is known. Then a GLS estimator for $\boldsymbol{\beta}$ is any $\mathbf{b}$ that minimizes

$$Q(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b})$$

The estimating equations are:

$$(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})\mathbf{b} = \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

and

$$\mathbf{b}_{GLS} = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^-(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y})$$

is a solution. For any estimable function $\mathbf{C}^T\boldsymbol{\beta}$, the unique b.l.u.e. is

$$\mathbf{C}^T\mathbf{b}_{GLS} = \mathbf{C}^T(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^-\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

with

$$V(\mathbf{C}^T\mathbf{b}_{GLS}) = \mathbf{C}^T(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^-\mathbf{C}$$

If $Y \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, then

$$\mathbf{C}^T\mathbf{b}_{GLS} \sim N\left(\mathbf{C}^T\boldsymbol{\beta},\ \mathbf{C}^T(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^-\mathbf{C}\right)$$

When $\mathbf{M}$ and/or $\mathbf{R}$ contain unknown parameters, you could obtain an "approximate BLUE" by replacing the unknown parameters with consistent estimators to obtain

$$\hat{\boldsymbol{\Sigma}} = \mathbf{Z}\hat{\mathbf{M}}\mathbf{Z}^T + \hat{\mathbf{R}}$$

and

$$\mathbf{C}^T\mathbf{b}^*_{GLS} = \mathbf{C}^T(\mathbf{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^-\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

- $\mathbf{C}^T\mathbf{b}^*_{GLS}$ is not a linear function of $\mathbf{y}$
- $\mathbf{C}^T\mathbf{b}^*_{GLS}$ is not a best linear unbiased estimator (BLUE)
- $\mathbf{C}^T(\mathbf{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^-\mathbf{C}$ tends to "underestimate" $V(\mathbf{C}^T\mathbf{b}^*_{GLS})$
- For "large" samples

$$\mathbf{C}^T\mathbf{b}^*_{GLS}\dot{\sim}N(\mathbf{C}^T\boldsymbol{\beta}, \mathbf{C}^T(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^-\mathbf{C})$$

6.4 Variance component estimation

- Estimation of parameters in $\mathbf{M}$ and $\mathbf{R}$
- Crucial to the estimation of estimable functions of fixed effects (e.g. $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$)
- Of interest in its own right (sources of variation in the pigment paste production example)

6.4.1 Rules for Expected Mean Squares

- **Rule 1.** The error term in the model, $e_{(ij\ldots)m}$, where the subscript m denotes the replication subscript.
- **Rule 2.** In addition to an μ and $e_{(ij\ldots)m}$, the model contains all the main effects and any interactions that the experimenter assumes exist.
- **Rule 3.** For each term in the model, divide the subscripts into three classes:
  1. live-those subscripts that are present in the term and are not in the parentheses;
  2. dead-those subscripts that are present in the term and are in parentheses; and
  3. absent-those subscripts that are present in the model but not in that particular term.

Thus, in $(\alpha\beta)_{ij}$, i and j are live and k is absent, and in $e_{(ij)k}$, k is live and i and j are dead.

- **Rule 4.** Degrees of freedom. The number of DF for any term in the model is the product of the number of levels associated with each dead subscript and the number of levels minus 1 associated with each live subscript. For example, the number of DF associated with $(\alpha\beta)_{ij}$ is $(a-1)(b-1)$, and the number of DF associated with $e_{(ij)k}$ is $ab(n-1)$.
- **Rule 5.** Each term in the model has either a variance component (random effect) or a fixed factor (fixed effect) associated with it. If an interaction contains at least one random effect, the entire interaction is considered random. A variance component has Greek letters as subscripts to identify the particular random effect. Thus, in a two-factor mixed model with factor A fixed and factor B random, the variance component for B is σ^2_b, and the variance component for AB is σ^2_{ab}. A fixed effect is always represented by the sum of squares of the model components associated with the factor divided by its degrees of freedom. For example, the effect of A is

$$\frac{\sum_{i=1}^{a}\alpha_i^2}{a-1}$$

.

- **Rule 6.** Expected mean squares. To obtain EMS, prepare the following table. There is a row for each model component (mean square) and a column for each subscript. Over each subscript, write the number of levels of the factor associated with that subscript and whether the factor is fixed (F) or random (R). Replicates are always considered to be random.
  1. In each row, write 1 if one of the dead subscripts in the row component matches the subscript in the column:

	F	F	R	
	a	b	n	
Factor	i	j	k	
α_i				
β_j				
$(\alpha\beta)_{ij}$				
$e_{(ij)k}$	1	1		

2. In each row, if any of the subscripts on the row component match the subscript in the column, write 0 if the column is headed by a fixed factor and 1 if the column is headed by a random factor:

	F	F	R	
	a	b	n	
Factor	i	j	k	
α_i	0			
β_j		0		
$(\alpha\beta)_{ij}$	0	0		
$e_{(ij)k}$	1	1	1	

3. In the remaining empty row positions, write the number of levels shown above the column heading:

	F	F	R	
	a	b	n	
Factor	i	j	k	
α_i	0	b	n	
β_j	a	0	n	
$(\alpha\beta)_{ij}$	0	0	n	
$e_{(ij)k}$	1	1	1	

4. To obtain the EMS for any model component. First cover all column headed by live subscripts on that component. Then, in each row that contains at least the same subscripts as those on the component being considered, take the product of the visible numbers and multiply by the appropriate fixed or random factor from rule 5. The sum of these quantities is the EMS of the model component being considered.

	F	F	R	
	a	b	n	
Factor	i	j	k	
α_i	0	b	n	$\sigma^2 + bn\sum \alpha_i^2/(a-1)$
β_j	a	0	n	$\sigma^2 + an\sum \beta_i^2/(b-1)$
$(\alpha\beta)_{ij}$	0	0	n	$\sigma^2 + \frac{n\sum\sum(\alpha\beta)_{ij}^2}{(a-1)(b-1)}$
$e_{(ij)k}$	1	1	1	σ^2

Example 4.
Use the rules for expected mean squares to derive the expected mean squares for the following models, and propose appropriate test statistics for all effects:

(a) $y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_{ijk}$

$$i = 1, 2, \ldots, a$$
$$j = 1, 2, \ldots, b$$
$$k = 1, 2, \ldots, n$$

Assuming that all the factors are random.

(b) $y_{ijk} = \mu + \alpha_i + b_j + (ab)_{ij} + e_{ijk}$

$$i = 1, 2, \ldots, a$$
$$j = 1, 2, \ldots, b$$
$$k = 1, 2, \ldots, n$$

Assuming that factor A is fixed and facor B is random.

6.4.2 ANOVA method (Method of Moments)

- Compute an ANOVA table
- Equate mean squares to their expected values
- Solve the resulting equations

Example 5. Penicillin production

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

where $\boldsymbol{\beta}_j \sim NID(0, \sigma^2_\beta)$ and $e_{ij} \sim NID(0, \sigma^2_e)$

Source of Variation	d.f.	Sums of Squares
Blocks	4	$a\sum_{j=1}^{b}(\bar{Y}_{.j} - \bar{Y}_{..})^2 = SS_{blocks}$
Processes	3	$b\sum_{i=1}^{a}(\bar{Y}_{1.} - \bar{Y}_{..})^2 = SS_{processes}$
error	12	$\sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij} - \bar{Y}_{1.} - \bar{Y}_{.j} + \bar{Y}_{..})^2 = SSE$
C. total	19	$\sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij} - \bar{Y}_{..})^2$

Start at the bottom:

$$MS_{error} = \frac{SSE}{(a-1)(b-1)}$$

$$E(MS_{error}) = \sigma^2_e$$

Then an unbiased estimator for σ_e is

$$\hat{\sigma}^2_e = MS_{error}$$

Next, consider the mean square for the random block effects:

$$MS_{blocks} = \frac{SS_{blocks}}{b-1}$$

$$E(MS_{blocks}) = \sigma^2_e + a\sigma^2_\beta$$

Then,

$$\sigma^2_\beta = \frac{E(MS_{blocks}) - \sigma^2_e}{a}$$

$$= \frac{E(MS_{blocks}) - E(MS_{error})}{a}$$

An unbiased estimator for σ^2_β is

$$\hat{\sigma}^2_\beta = \frac{MS_{blocks} - MS_{error}}{a}$$

For the penicillin data

$\hat{\sigma}^2_e =$

$\hat{\sigma}_\beta =$

$\widehat{V(Y_{ij})} =$

6.4.3 Properties of ANOVA methods for variance component estimation

(i) Broad applicability

- easy to compute in balanced cases
- ANOVA is widely known
- not required to completely specify distributions for random effects

(ii) Unbiased estimators

(iii) Sampling distribution is not exactly known, even under the usual normality assumptions (except for $\hat{\sigma}^2_e = MS_{error}$)

(iv) May produce negative estimates of variances

(vi) For unbalanced studies, there may be no "natural" way to choose

$$\hat{\sigma}^2 = \sum_{i=1}^{k} a_i MS_i$$

Result 1. If $MS_1, MS_2, \ldots, MS_k$ are distributed independently with

$$\frac{(df_i)MS_i}{E(MS_i)} \sim \chi^2_{df_i}$$

and constants $a_i > 0, \quad i = 1, 2, \ldots, k$ are selected so that

$$\hat{\sigma}^2 = \sum_{i=1}^{k} a_i MS_i$$

has expectation σ^2, then

$$V(\hat{\sigma}^2) = 2\sum_{i=1}^{k} \frac{a_i^2[E(MS_i)]^2}{df_i}$$

and an unbiased estimator of this variance is

$$\widehat{V}(\hat{\sigma}^2) = \frac{2\sum a_i^2 MS_i^2}{(df_i + 2)}$$

Furthermore,

$$\begin{aligned} E(MS_i^2) &= V(MS_i) + [E(MS_i)]^2 \\ &= \frac{2[E(MS_i)]^2}{df_i} + [E(MS_i)]^2 \\ &= \left(\frac{df_i + 2}{df_i}\right)[E(MS_i)]^2 \end{aligned}$$

Consequently,

$$E\left[2\sum_{i=1}^{k} \frac{a_i^2 MS_i^2}{(df_i + 2)}\right] = V(\hat{\sigma}^2)$$

A "standard error" for

$$\hat{\sigma}^2 = \sum_{i=1}^{k} a_i MS_i$$

could be reported as

$$S_{\hat{\sigma}^2} = \sqrt{2\sum_{i=1}^{k} \frac{a_i^2 MS_i^2}{(df_i + 2)}}$$

Using the Cochran-Satterthwaite approximation, an approximate $(1-\alpha) \times 100\%$ confidence interval for σ^2 could be constructed as:

$$\begin{aligned} 1 - \alpha &\doteq Pr\left\{\chi^2_{\nu,1-\alpha/2} \leq \frac{v\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{\nu,\alpha/2}\right\} \\ &= Pr\left\{\frac{v\hat{\sigma}^2}{\chi^2_{\nu,\alpha/2}} \leq \sigma^2 \leq \frac{v\hat{\sigma}^2}{\chi_{\nu,1-\alpha/2}}\right\} \end{aligned}$$

where $\hat{\sigma}^2 = \sum_{i=1}^{k} a_i MS_i$ and

$$v = \frac{\left[\sum_{i=1}^{k} a_i MS_i\right]^2}{\sum_{i=1}^{k} \frac{[a_i MS_i]^2}{df_i}}$$

Example 6. Pigment production
In this example the main objective is the estimation of the variance components

Source of Variation	d.f.	MS	E(MS)
Batches	15-1=14	86.495	$\sigma_e^2 + 2\sigma_\delta^2 + 4\sigma_\beta^2$
Samples in Batches	15(2-1)=15	57.983	$\sigma_e^2 + 2\sigma_\delta^2$
Tests in Samples	(30)(2-1)=30	0.917	σ_e^2

(a) Estimates of variance components.

(b) Find a 95% confidence interval for μ.

Example 7.

A study was conducted on human subjects to measure the effects of 3 foods on serum glucose levels. Each of the 3 foods was randomly assigned to 5 subjects. The serum glucose was measured for each of the subjects at 7 different time points starting at 15 minutes and every 15 minutes after food was ingested. Consider the model

$$Y_{ijk} = \mu + \alpha_i + S_{ij} + \tau_k + \gamma_{ik} + e_{ijk}$$

where y_{ijk} is the serum glucose levels at the k^{th} time point for the j^{th} subject with the i^{th} food, α_i is the fixed diet effect, τ_k is the fixed time effect and γ_{ik} is the fixed diet $\times$ time effect, $S_{ij} \sim NID(0, \sigma_S^2)$ and is independent of $e_{ijk} \sim NID(0, \sigma_e^2)$.

(a) Find $V(\mathbf{Y_{ij}})$, for this model?

(b) Provide the formulas for the estimator of σ_e^2 and σ_S^2.

(c) What is the correlation between observations taken on the same subject?

(d) Find the estimator of $V(\bar{Y}_{ij.})$ and provide it's degrees of freedom.

(e) Find the estimator of $V(\bar{Y}_{i.k})$ and provide it's Satterthwaith degrees of freedom.

Example 8.

An experiment was conducted to compare the effectiveness of two diets (denoted D_1 and D_2). The subjects included 60 men between the ages of 18 to 35. Each subject lifts the dumbbells until his muscles were depleted of energy, rested for two hours, and lifts the dumbbells again until exhaustion. During the rest period, each subject eat one of the two diets as assigned by the researchers. Each subject's performance on the second round of lifting dumbbells following the rest period was assigned a score between 0 and 100 based on the energy expended prior to exhaustion. Higher scores indicate of better performance. 20 of the 60 subjects repeated the lift rest lift trial on a second occasion separated from the first by approximately three weeks. These subjects eat one diet during the first trial and the other during the second trial. The diet order was randomized for each subject by the researchers. The other 40 subjects performed the trial only a single time, eating a randomly assigned diet during the rest period. 20.0 of these subjects received diet 1, and the other 20.0 received diet 2. A portion of the entire data set is provided in the following table.

Subject	Diet 1	Diet 2
1	40	50
2	35	38
⋮	⋮	⋮
20	28	55
21	35	–
22	80	-
⋮	⋮	⋮
40.0	55	-
41.0	-	17
42.0	-	54
⋮	⋮	⋮
60	-	60

Subjects 1 through 20 in the table above represent the 20 subjects who performed the trial separately for each of the diets. Note that the data set contains no information about which diet was received in the first trial and which drink was received in the second trial. Suppose the following model is appropriate for the data.

$$y_{ij} = \mu_i + u_j + e_{ij}, (1)$$

where y_{ij} is the score for diet i and subject j μ_i is the unknown mean score for diet i, u_j is a random effect corresponding to subject j, and e_{ij} is a random error corresponding to the score for diet i and subject j ($i = 1, 2$ and $j = 1, 2, \ldots, 60$). Here $u_1, u_2, \ldots, u_{60}$ are assumed to

be independent and identically distributed as $N(0, \sigma_u^2)$ and independent of the e_{ij}'s, which are assumed to be independent and identically distributed as $N(0, \sigma_e^2)$.

(a) For each of the subjects who received both diets, the difference between the scores ($D_1 - D_2$ score) was computed. This yielded 20 score differences denoted $d_1, d_2, \ldots, d_{20}$ Describe the distribution of these differences considering the assumptions about the distribution of the original scores in model (1).

(b) Suppose you were given only the differences from part (a). Provide a formula for a test statistic (as a function of $d_1, d_2, \ldots, d_{20}$) that could be used to test $H_0 : \mu_1 - \mu_2$.

(c) Let $a_1, a_2, \ldots, a_{20.0}$ be the scores of the subjects who received only diet 1. Let $b_1, b_2, \ldots, b_{20.0}$ be the scores of the subjects who received only diet 2. Suppose you were given only these 20.0 scores. Determine the BLUE of $\mu_1 - \mu_2$ and the variance of this estimator.

6.5 Estimation of fixed effects

Denote the resulting REML estimators as

$$\hat{\mathbf{M}} \quad \hat{\mathbf{R}} \quad \text{and} \quad \hat{\mathbf{\Sigma}} = \mathbf{Z}\hat{\mathbf{M}}\mathbf{Z}^T + \hat{\mathbf{R}}$$

For any estimable function $\mathbf{C}\boldsymbol{\beta}$, the **blue** is the generalized least squares estimator

$$\mathbf{C}\mathbf{b}_{GLS} = \mathbf{C}(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{y}$$

Using the REML estimator for

$$\mathbf{\Sigma} = \mathbf{Z}\mathbf{M}\mathbf{Z}^T + R$$

an approximation is

$$\mathbf{C}\hat{\boldsymbol{\beta}} = \mathbf{C}(\mathbf{X}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^-\mathbf{X}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{y}$$

and for "large" samples:

$$\mathbf{C}\hat{\boldsymbol{\beta}} \sim N(\mathbf{C}\boldsymbol{\beta}, \mathbf{C}(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^-\mathbf{C}^T)$$

6.6 Prediction of random effects

Given the observed responses $\mathbf{y}$, predict the value of $\mathbf{u}$. For our model,

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{M} & 0 \\ 0 & R \end{bmatrix} \right) .$$

Then

$$\begin{aligned} \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \end{bmatrix} &= \begin{bmatrix} \mathbf{u} \\ \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} \\ \mathbf{X}\boldsymbol{\beta} \end{bmatrix} + \begin{bmatrix} \mathbf{I} & 0 \\ \mathbf{Z} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} \\ &\sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{X}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \mathbf{M} & \mathbf{M}\mathbf{Z}^T \\ \mathbf{Z}\mathbf{M} & \mathbf{Z}\mathbf{M}\mathbf{Z}^T + R \end{bmatrix} \right) \end{aligned}$$

The Best Linear Unbiased Predictor (BLUP) is the b.l.u.e. for

$$\begin{aligned} &E(\mathbf{u}|\mathbf{y}) \\ &= E(\mathbf{u}) + (\mathbf{M}\mathbf{Z}^T)(\mathbf{Z}\mathbf{M}\mathbf{Z}^T + R)^{-1}(\mathbf{y} - E(\mathbf{y})) \\ &= \mathbf{0} + \mathbf{M}\mathbf{Z}^T(\mathbf{Z}\mathbf{M}\mathbf{Z}^T + R)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

Substitute the b.l.u.e. for $\mathbf{X}\boldsymbol{\beta}$,

$$\mathbf{X}\mathbf{b}_{GLS} = \mathbf{X}(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{y}$$

Then, the BLUP for $\mathbf{u}$ is

$$\begin{aligned} &BLUP(\mathbf{u}) \\ &= \mathbf{M}\mathbf{Z}^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b}_{GLS}) \\ &= \mathbf{M}\mathbf{Z}^T\mathbf{\Sigma}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{\Sigma}^{-1})\mathbf{y} \end{aligned}$$

When $\mathbf{M}$ and $\mathbf{\Sigma} = \mathbf{Z}\mathbf{M}\mathbf{Z}^T + R$ are known. Substituting REML estimators $\hat{\mathbf{M}}$ and $\hat{\mathbf{R}}$ for $\mathbf{M}$ and $\mathbf{R}$, an approximate BLUP for $\mathbf{u}$ is

$$\begin{aligned}\hat{\mathbf{u}} &= \hat{\mathbf{M}}\mathbf{Z}^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^-\mathbf{X}^T\hat{\mathbf{\Sigma}}^{-1})\mathbf{y} \\ &= \hat{\mathbf{M}}\mathbf{Z}^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{y} - \underline{\mathbf{X}\hat{\boldsymbol{\beta}}})\end{aligned}$$

For "large" samples, the distribution of $\hat{\mathbf{u}}$ is approximately multivariate normal with mean vector $\mathbf{0}$ and covariance matrix

$$\mathbf{M}\mathbf{Z}^T\mathbf{\Sigma}^{-1}(\mathbf{I} - P)\mathbf{\Sigma}(\mathbf{I} - P)\mathbf{\Sigma}^{-1}\mathbf{Z}\mathbf{M}$$

where

$$P = \mathbf{X}(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{\Sigma}^{-1}$$

Given estimates $\hat{\mathbf{M}}, \hat{\mathbf{R}}$ and $\hat{\mathbf{\Sigma}} = \mathbf{Z}\hat{\mathbf{M}}\mathbf{Z}^T + \hat{\mathbf{R}}$,
$\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ provide a solution to the mixed model equations:

$$\begin{bmatrix}\mathbf{X}^T\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}^T\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\hat{\mathbf{R}}^{-1} & \mathbf{Z}^T\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{M}}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^T\hat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{Z}^T\hat{\mathbf{R}}^{-1}\mathbf{y}\end{bmatrix}$$

A generalized inverse of

$$\begin{bmatrix}\mathbf{X}^T\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}^T\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\hat{\mathbf{R}}^{-1} & \mathbf{Z}^T\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{M}}^{-1}\end{bmatrix}$$

is used to approximate the covariance matrix for $\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}\end{bmatrix}$

Example 9.

Suppose 2 maize genotypes were assigned to 6 plots in a field using an unbalanced completely randomized design. Plots were planted with seed from their assigned genotypes(Genotype 1 and 2 were assigned to 4 and 2 plots respectively), and yield in bushels per acre was recorded for each plot at the end of the growing season. Consider the model

$$y_{ij} = \mu + g_i + e_{ij},$$

where $\mu + g_i$ is the mean yield for the i^{th} genotype, and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$ for all i and j. Assume $g_1, g_2 \overset{iid}{\sim} N(0, \sigma_g^2)$ and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$. Find the BLUP of $\mu + g_1$.